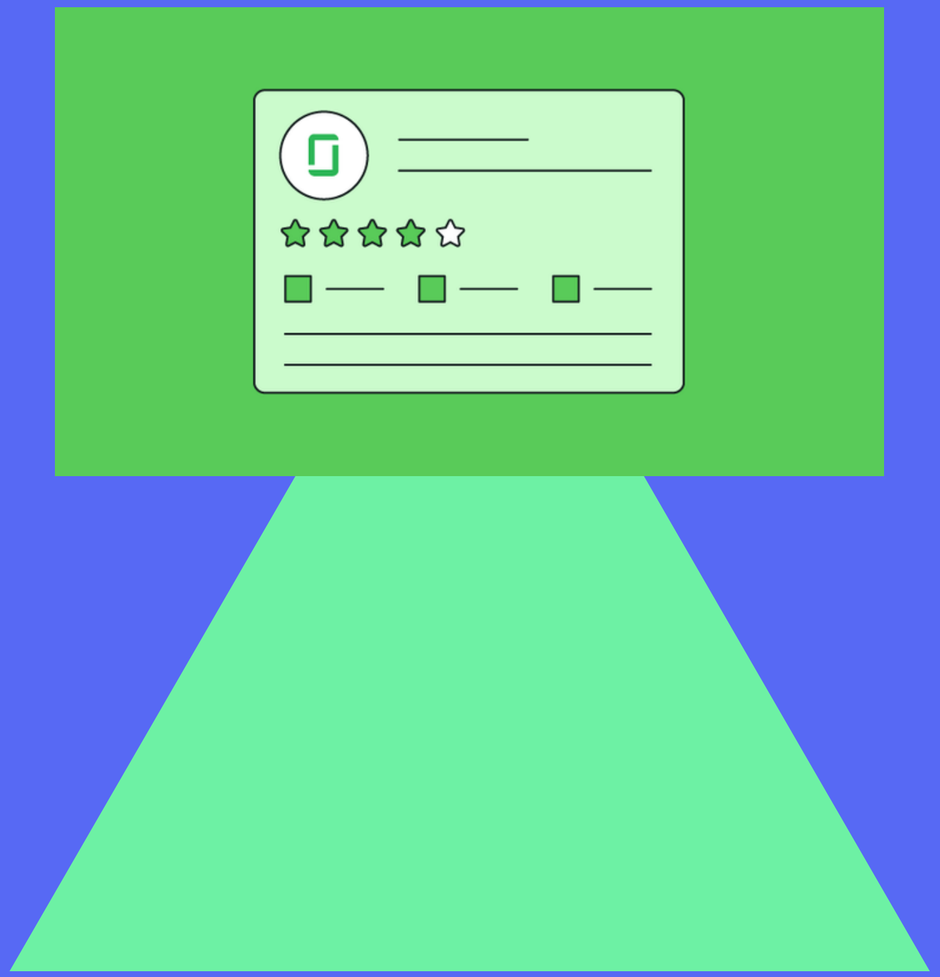
Walking the Walk

Turning Qualitative Text into Quantitative Results



OVERVIEW

The process of finding a job is quite an undertaking with many factors to consider. Reviews from current and former employees can play a key role in whether or not a candidate chooses to apply. In this project, I attempt to translate qualitative reviews into a quantitative metric which will allow candidates the ability to trust the simple ratings to a greater degree as they will be more reflective of the reviewer's true impression of the company. This may also help companies better understand what their trouble areas are since they may be more nuanced than what the ratings currently indicate.

Summary of Approach

The 4 Phases

DATA ACQUISITION

Web scraped 14k+ reviews from 150 companies of various industries and ratings from the Glassdoor site.

EDA

During this step non-text and text features was cleaned, recategorized, and visualized..

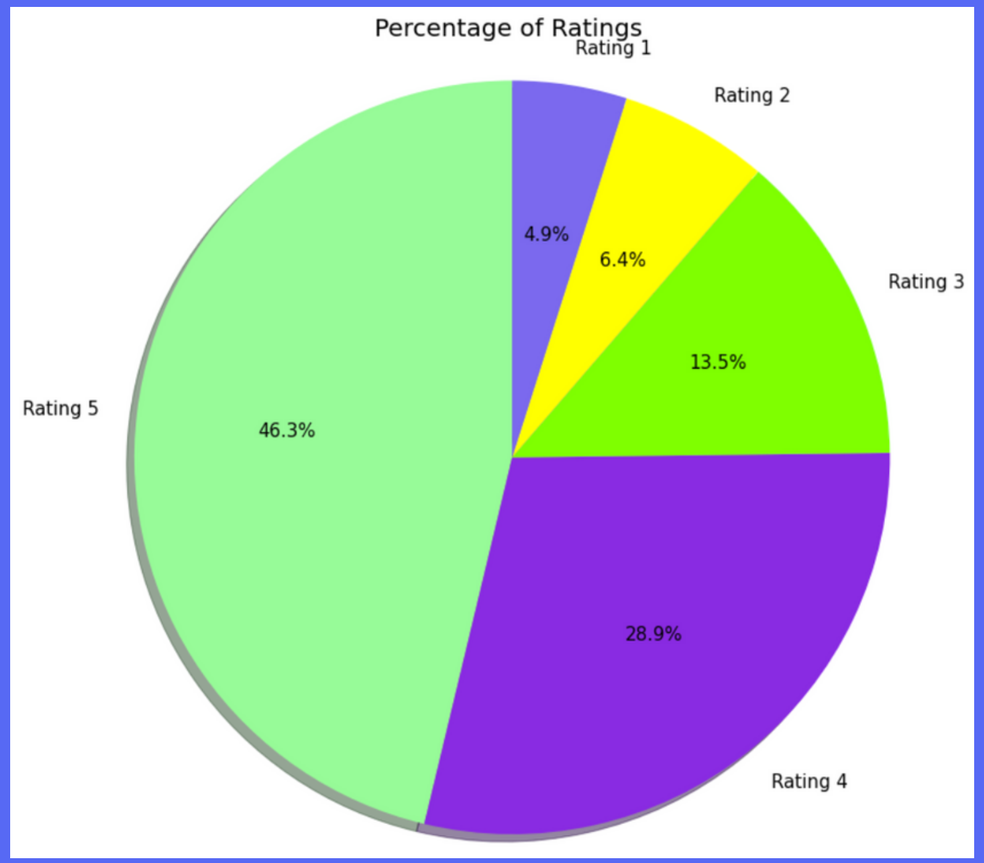
PROCESSING

Using the insights from EDA the data was then process and prepared for modeling the actions such as tokenization, removing stop words, and vectorizing.

MODEL

Laslty, I tested several models in order to find the best fit for my data.

Looking At the Target



Thing Don't Add Up

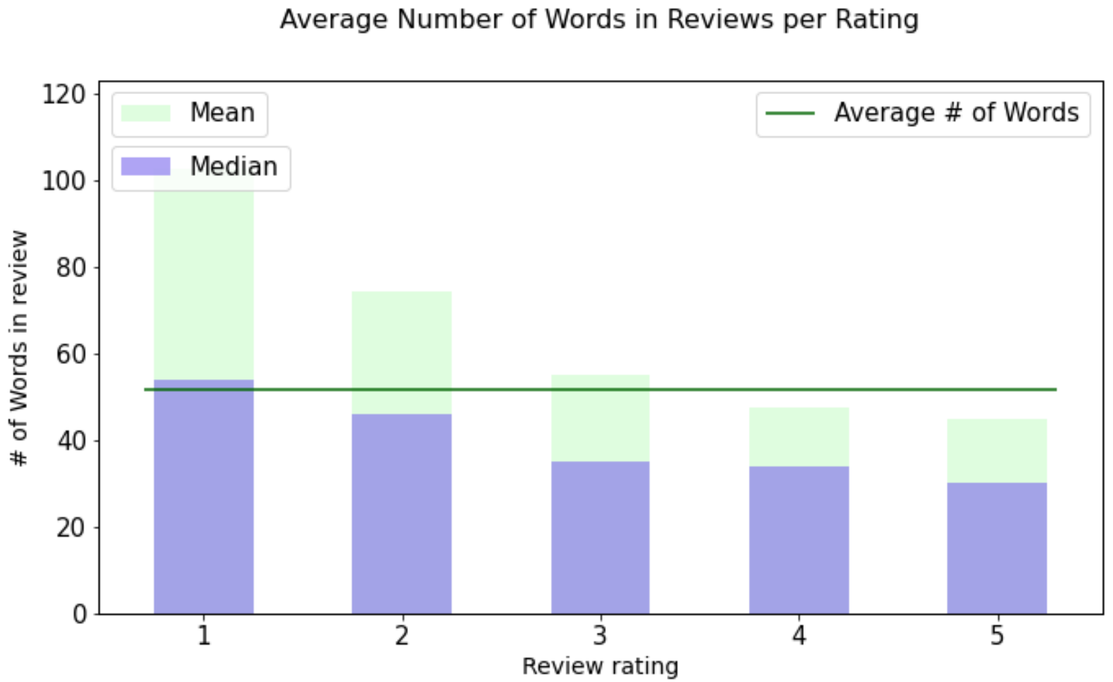
When looking at how users rated the individual aspects of their company they had little to do with the overall score given

Top 100 Most Common Words

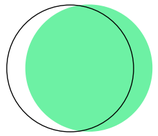
We can see that 'helpfu' is the most popular word throughout all the reviews. We can surmised that reviewers had a lot to say about what they did and did not find helpful at the places of work.

The Worse The Rating...

Another interesting find was that the lower rating the larger the average word count. The worse the experience the more there was to say about it.

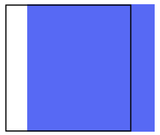


Model Tests



RANDOM FOREST

Random forest performed very near my best performer with 58% accuracy.



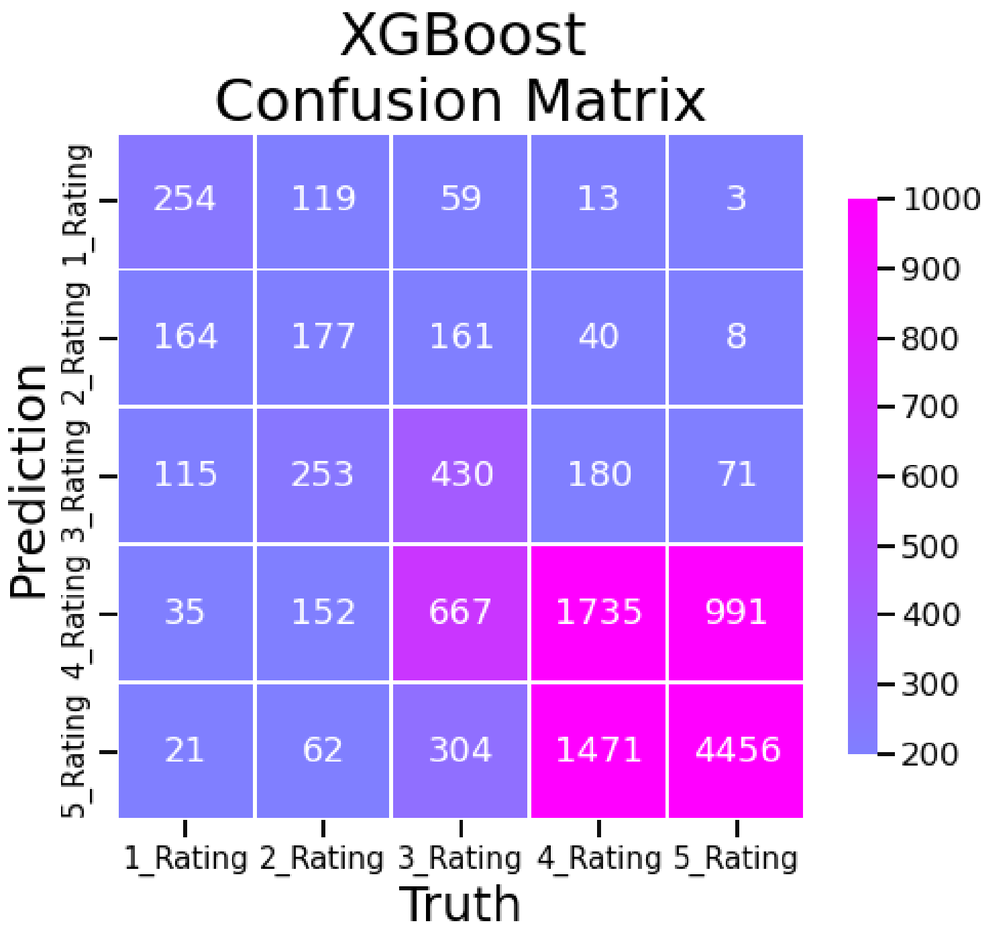
KNN

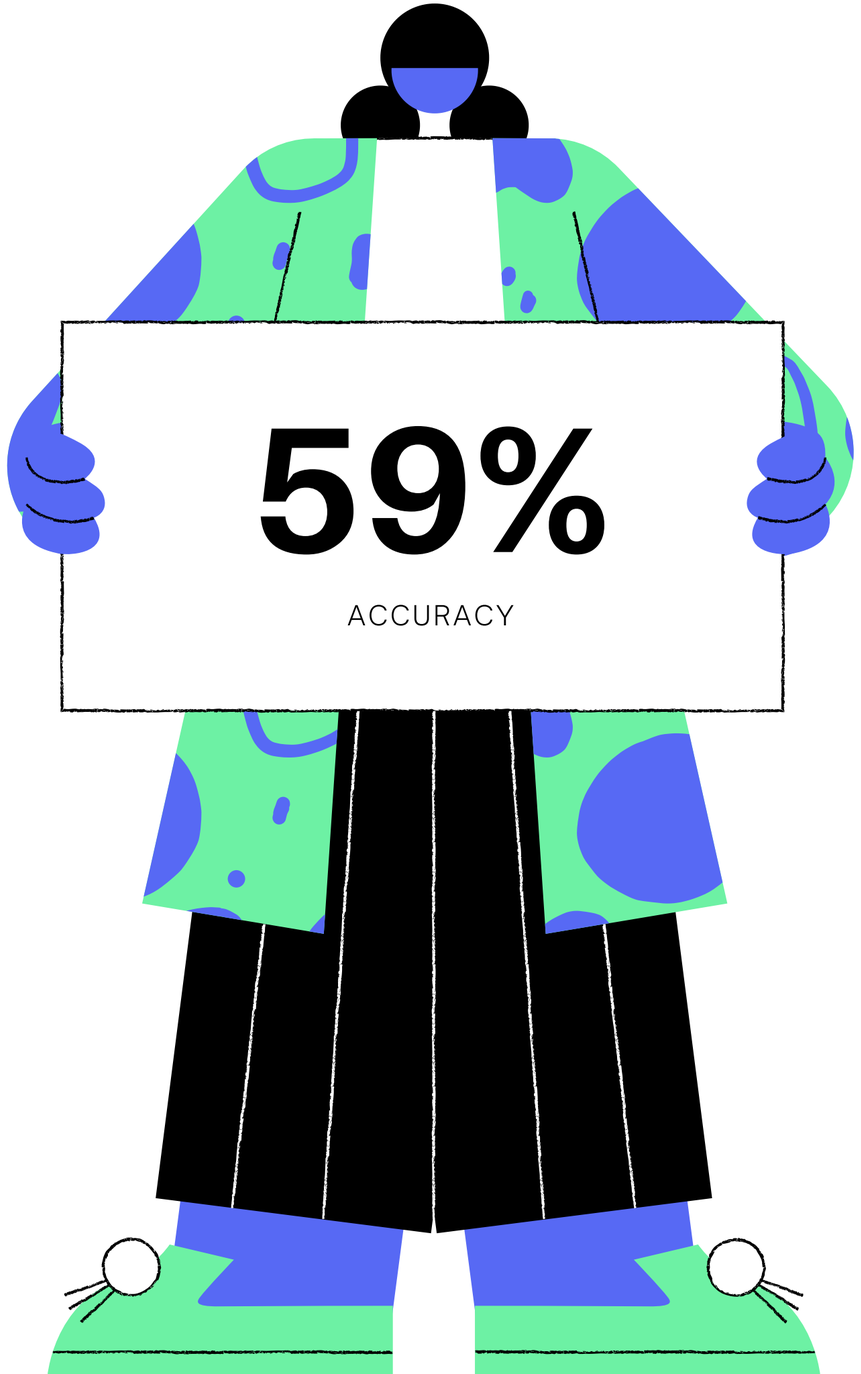
The KNN Model was the worst performer resulting in an only 45% accuracy.

LIGHTGBM

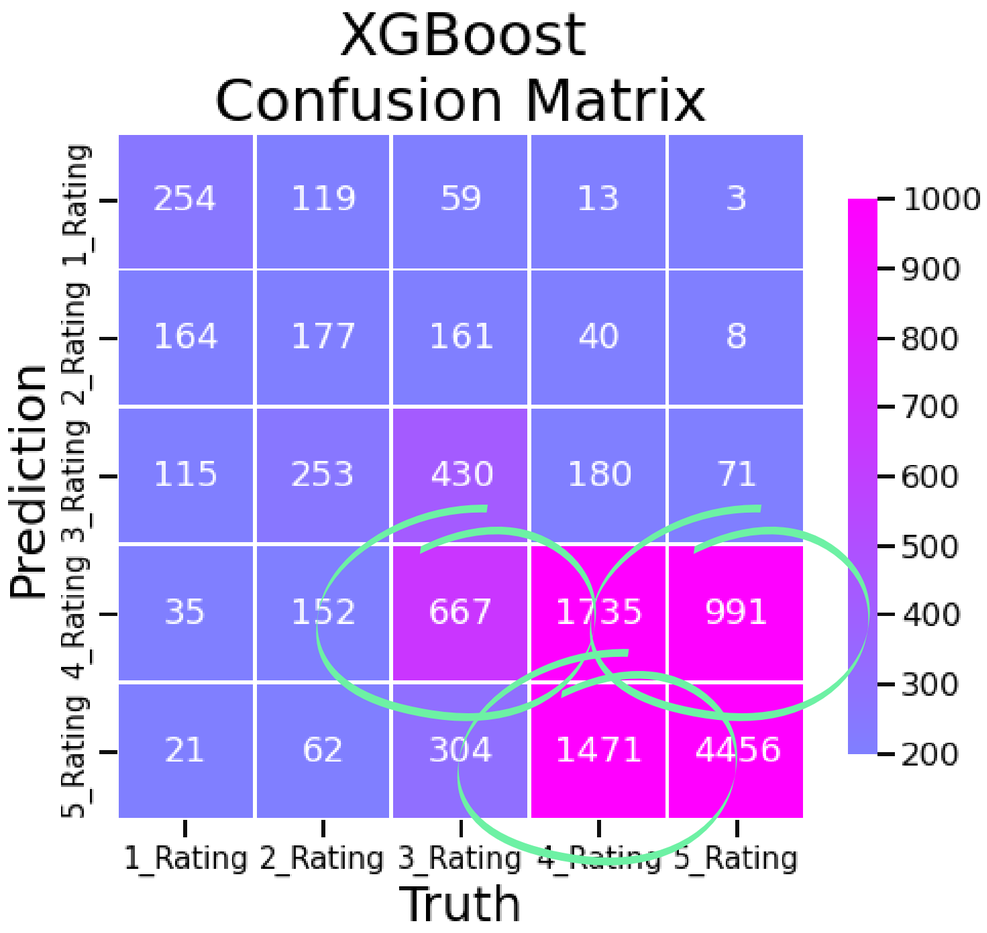
Light Gradient Boosted Machine also resulted in a 58% accuracy.

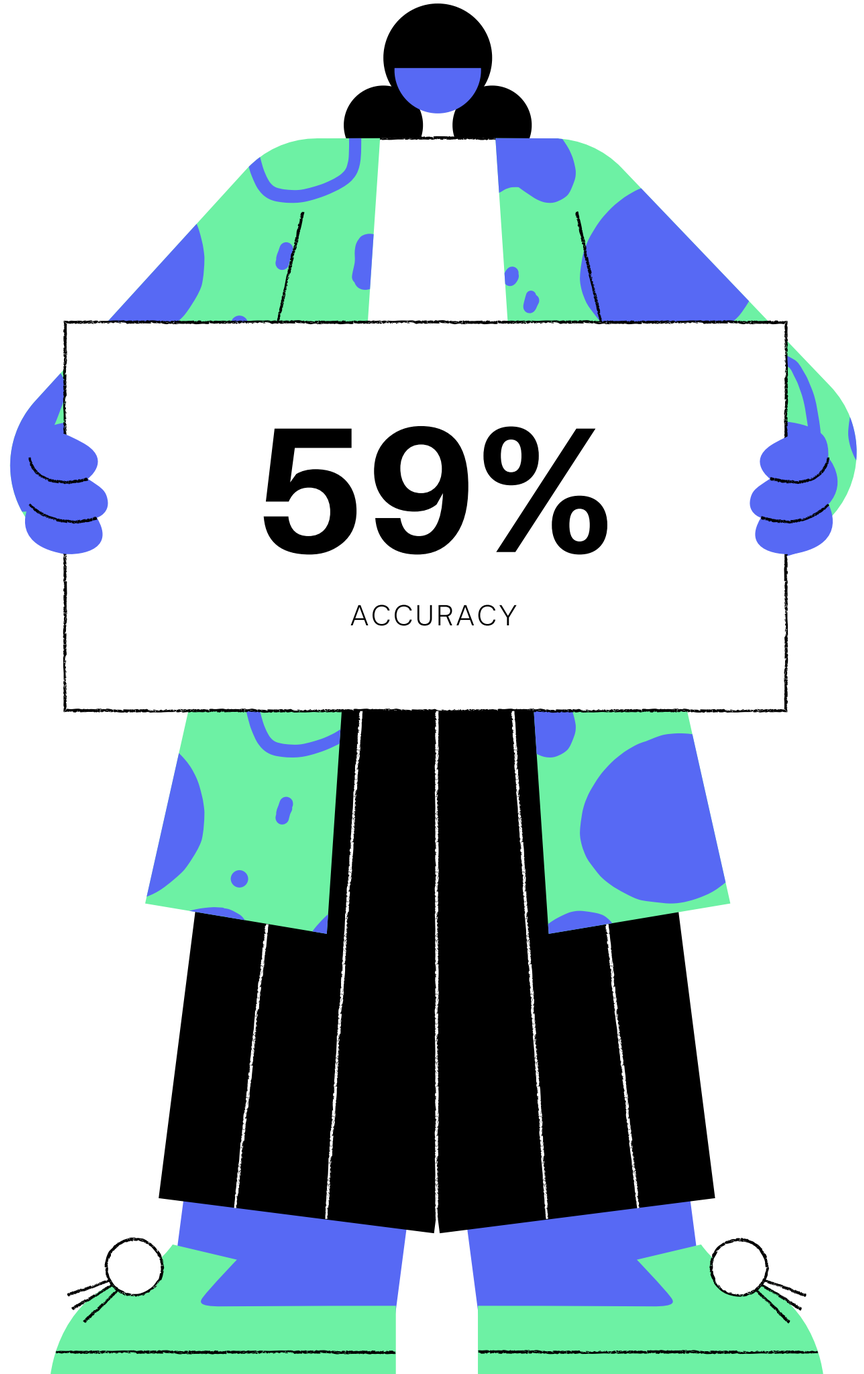
XGBOOST MODEL





XGBOOST MODEL





Conclusion

This indicates that machine learning could be a better way to score company reviews and this the overall company rating versus relying on a self reported score.

Next Step

In future iterations of this project I hope to increase my sample and move try out deep learning models to improve results.

Contact Me

I'd love to connect.

LINKEDIN

https://www.linkedin.com/in/anjcray/

EMAIL ADDRESS

dataonatangent@gmail.com

GITHUB

https://github.com/DataOnATangent